Signal Identification for Rare and Weak Features: Higher Criticism or False Discovery Rates?

Bernd Klaus *and Korbinian Strimmer *

12 December 2011; last revised 19 July 2012

Abstract

Signal identification in large-dimensional settings is a challenging problem in biostatistics. Recently, the method of higher criticism (HC) was shown to be an effective means for determining appropriate decision thresholds. Here, we study HC from a false discovery rate (FDR) perspective. We show that the HC threshold may be viewed as an approximation to a natural class boundary (CB) in two-class discriminant analysis which in turn is expressible as FDR threshold. We demonstrate that in a rare-weak setting in the region of the phase space where signal identification is possible both thresholds are practicably indistinguishable, and thus HC thresholding is identical to using a simple local FDR cutoff. The relationship of the HC and CB thresholds and their properties are investigated both analytically and by simulations, and are further compared by application to four cancer gene expression data sets.

*Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig, Härtelstr. 16–18, D-04107 Leipzig, Germany

1

1 Introduction

Identification of sparse and weak signals in complex high-dimensional data is a challenging statistical problem that has many important applications in fields as diverse as astronomy, finance, genetics, medicine, and proteomics. A typical biomedical task is the search for biomarkers using data from genome-wide association studies (Xie et al., 2011). Signal *identification* is much more difficult than the closely related problem of signal *detection*. Whereas in detection we are concerned purely with the presence or absence of a signal, in identification we additionally seek to locate the signal.

In a series of recent publications the method of "Higher Criticism" (HC) was powerfully advocated in settings with rare and weak features as an efficient means for signal detection (Donoho and Jin, 2004) as well as signal identification (Donoho and Jin, 2008, 2009). Originally, HC was introduced by Tukey (1976) as an approach to multiple significance testing using a second level test statistic computed from p-values. Importantly, in Donoho and Jin (2004) it was shown that HC provides a procedure that is optimal for signal detection in the sense that it achieves the best possible theoretical detection limit discovered earlier by Ingster (1999). Subsequently, HC was also employed in a thresholding procedure to determine relevant features for prediction. Again, it was demonstrated that the HC approach to signal identification outperforms other commonly employed selection strategies, in particular those based on false discovery rates (Donoho and Jin, 2008, 2009).

In Ahdesmäki and Strimmer (2010) the utility of HC for variable selection in classification was confirmed but at the same time it was also empirically shown that in the signal identification problem controlling the false *non*-discovery rate is equivalent to the HC procedure. Furthermore, it was discovered by Jager and Wellner (2007) that HC is not unique in achieving the detection limit. Given the success of HC this raises questions about the fundamental principles that may underlie this approach.

Here, we explore signal identification using the HC and false (non)-discovery approaches, with the aim to provide a better understanding of HC as well as offering a simple explanation for HC's favorable performance. Specifically, we argue that the decision threshold provided by HC may also be viewed as an approximation to a natural class boundary (CB) in classification which in turn is easy to understand from a false discovery rate perspective. In particular, in the rare-weak setting in the region of the phase space where identification is actually possible we show that the HC and CB threshold are nearly indistinguishable.

The remainder of the paper is structured as follows. First, we provide a non-technical introduction to HC both on sample and population level. Second, we derive the ideal thresholds corresponding to HC and false discovery rate approaches, and explore their mutual relationships. Next, we investigate these thresholds in the rare-weak model and establish the near identity of HC and a natural CB threshold in the rare-weak identification setting. Finally, we demonstrate the validity of the theoretical considerations by simulation and by analyzing data from four gene expression experiments.

2 Higher Criticism

In the following, we introduce the HC approach to signal identification, and discuss various properties of the HC threshold both from a sample and population point of view.

2.1 Empirical HC threshold based on p-values

We consider a situation with d observed test statistics $y_1, \ldots, y_d$. For each statistic we compute a corresponding p-value $p_1, \ldots, p_d$. The dimension d is potentially very large, as in many current applications in genomics or proteomics.

The HC approach to signal identification then proceeds as follows:

- First, by arranging the p-values from smallest to largest $p_{(1)}, \ldots, p_{(d)}$ the empirical distribution function of the p-values is obtained,

$$\hat{F}(x) = i/d \text{ for } p_{(i)} \leq x < p_{(i+1)}$$

  with $x \in [0;1]$, $p_{(0)} = 0$ and $p_{(d+1)} = 1$.

- Second, the empirical HC objective function

$$\widehat{HC}(x) = \frac{|\hat{F}(x) - x|}{\sqrt{\hat{F}(x)(1 - \hat{F}(x))/d}} \tag{1}$$

  is computed (Donoho and Jin, 2008, 2009).

- Third, the HC statistic $\widehat{HC}^\star$ is obtained as the maximum of the empirical HC objective

$$\widehat{HC}^\star = \max_i \widehat{HC}(p_{(i)}) = \widehat{HC}(x^{\text{HC}}) \,.$$

- Finally, the maximizing argument x^{HC} is taken as the HC decision threshold for signal identification. As shown in Fig. **1a** all $p_i < x^{\text{HC}}$ are considered "significant" and thus assumed to likely correspond to non-null cases.

Informally, the empirical HC objective function $\widehat{HC}(x)$ may be interpreted as z-scores constructed from p-values — recall that $\text{Var}(\hat{F}(x)) = F(x)(1 - F(x))/d$. Indeed, it is precisely this second level assessment of p-values that was the original motivation for the HC approach (Tukey, 1976) and that gave rise to its name "Higher Criticism".

2.2 Population HC objective function and goodness-of-fit statistics

By definition, p-values have a uniform $U(0,1)$ null distribution with $F_0(x) = x$. Moreover, the marginal distribution of the p-values may be viewed as a two-component mixture

$$F(x) = \eta_0 F_0(x) + (1 - \eta_0)F_A(x)$$

3
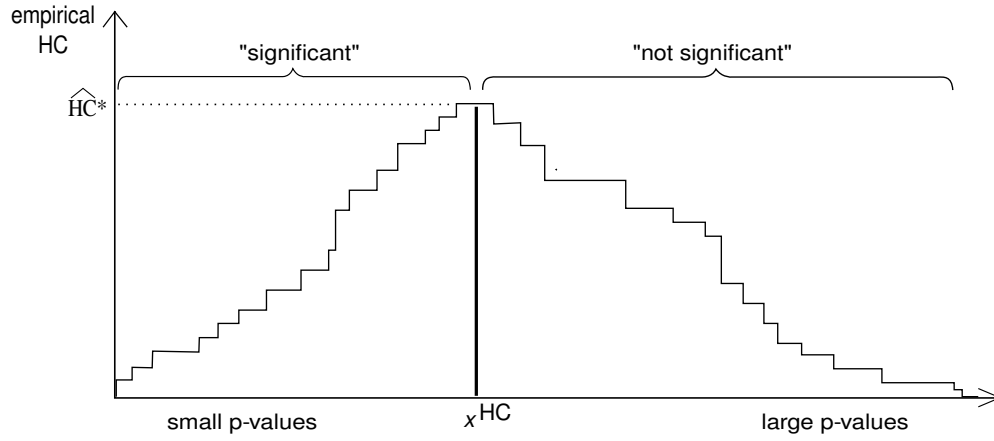
a) HC Threshold

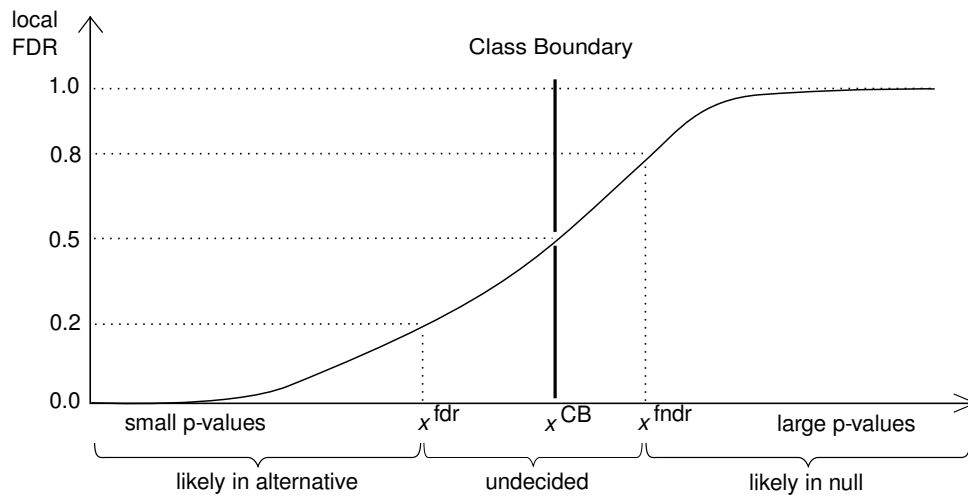

b) CB and FDR Thresholds

Figure 1: a) Empirical HC decision threshold x^{HC} obtained by maximizing the empirical HC objective, and b) Class boundary x^{CB} given by local FDR = $1/2$ and its relationship to the neighboring local FDR and local FNDR thresholds.

Table 1: Relationship of HC statistic with other goodness-of-fit statistics.

	Supremum	Expectation
Not standardized	Kolmogorov-Smirnov: $\sup_x \|F_A(x) - F_0(x)\|$	Cramér-von Mises: $E_F\{(F_A(X) - F_0(X))^2\}$
Standardized	Higher Criticism: $\sup_x \left\{ \frac{\|F_A(x) - F_0(x)\|}{\sqrt{F(x)(1-F(x))}} \right\}$	Anderson-Darling: $E_F \left\{ \frac{(F_A(X) - F_0(X))^2}{F(X)(1-F(X))} \right\}$

of the null model $F_0(x)$ and an alternative model $F_A(x)$ where $\eta_0 \in [0; 1]$ is the proportion of the null. With this in mind the squared empirical HC objective function can be written as

$$\widehat{HC}(x)^2 \propto \frac{(\hat{F}_A(x) - F_0(x))^2}{\hat{F}(x)(1 - \hat{F}(x))} \, .$$

The proportionality factor $d(1 - \eta_0)^2$ has been left out as it does not depend on x and hence is irrelevant for determining the decision threshold x^{HC}. Thus, for maximization we can use the above formula rather than Eq. **1**. Furthermore, it has the advantage of immediately generalizing to the *population level* (i.e. to $d \to \infty$)

$$HC(x)^2 \propto \frac{(F_A(x) - F_0(x))^2}{F(x)(1 - F(x))} \tag{2}$$

which greatly facilitates the conceptual understanding of the HC approach.

The function Eq. **2** is well known from the goodness-of-fit statistic of Anderson and Darling (1954) which is proportional to the expectation $E_F(HC(X)^2)$. Hence, the HC statistic bears the same relationship to the Anderson-Darling statistic as does the Kolmogorov-Smirnov statistic to the Cramér-van Mises statistic (Darling, 1957). Moreover, as can be seen in Tab. **1** the HC statistic is the standardized Kolmogorov-Smirnov (KS) statistic. In fact, the KS statistic may used in the same fashion as HC to derive a decision threshold x^{KS}.

In the mixture model for p-values it is commonly assumed (see also Section 3 on false discovery rates) that $F_A(x) \geq F_0(x)$ for all x, i.e. that the the alternative component is stochastically smaller than or equal to the null component. Thus, on population level (though not on sample level) we may leave out the absolute value signs in the first column of Tab. **1**.

2.3 Invariance of HC objective function

By inspection of Eq. **2** we can derive a number of interesting properties of the HC objective function.

First, it is completely symmetric with regard to the two components in the underlying mixture model for the p-values. The alternative model F_A and the null model F_0 play the same role in Eq. **2**.

Second, for computing the HC objective it is not necessary to explicitly specify the null proportion η_0.

Third, Eq. **2** is invariant against transformation of the underlying test statistic. This can be seen as follows: Under a change of variables from x to $y = y(x)$ the distribution function changes according to

$$F^Y(y) = \begin{cases} F\left(x(y)\right) & \text{for increasing } x(y), \text{ and} \\ 1 - F\left(x(y)\right) & \text{for a decreasing transformation.} \end{cases}$$

Applied to Eq. **2** this leads to

$$HC(y)^2 = HC(y(x))^2 \propto \frac{(F_A^Y(y) - F_0^Y(y))^2}{F^Y(y)(1 - F^Y(y))} \, .$$

Remarkably, the HC objective function Eq. **2** retains its functional form under a change of variables. Thus, Eq. **2** is *not* constrained to p-values only and may instead be applied to any test statistic y without the need of prior conversion to the p-value scale. The HC decision threshold as the location of the maximum of Eq. **2** transforms accordingly, from x^{HC} to $y^{\text{HC}} = y(x^{\text{HC}})$.

3 False Discovery Rates

For comparison we now briefly recapitulate the "False Discovery Rate" (FDR) approach to signal identification. Like the HC approach it is also best understood on the population level. For comprehensive overview see, e.g., Efron (2008).

3.1 Definition of FDR and related quantities

Essentially, there are two variants of FDR criteria, one based on distributions (tail area-based FDR) and the other on densities (local FDR). In addition, if the roles of null and alternative are interchanged one arrives at the "False Non-Discovery Rate" (FNDR).

On the p-value scale, the tail-area-based FDR (or Fdr) is defined as

$$\text{Fdr}(x) = \text{Pr}(\text{"null"}|X \le x) = \frac{\eta_0 F_0(x)}{F(x)} = \frac{\eta_0 x}{F(x)} \, .$$

By construction, $\text{Fdr}(x)$ is the proportion of p-values from the null component found among all p-values smaller than x. In order for $\text{Fdr}(x)$ to be monotonically increasing with x (i.e. to ensure that the ordering of test statistics does not change) it is necessary that $f_A(x)$ is a monotonically decreasing density, and thus both $F_A(x)$ and $F(x)$ must be

6

assumed to be concave (Langaas et al., 2005; Strimmer, 2008b). This also implies that the alternative and null are stochastically ordered with $F_A(x) \geq F_0(x)$ for all x. The empirical estimate of Fdr for a set ordered p-values $p_{(1)}, \ldots, p_{(d)}$ is the rule of Benjamini and Hochberg (1995),

$$\widehat{\text{Fdr}}(p_{(i)}) = \frac{\hat{\eta}_0 p_{(i)}}{\hat{F}(p_{(i)})} \leq \frac{d}{i} p_{(i)},$$

which also shows that Fdr may be viewed as a multiplicity-adjusted p-value. As complementary error one also studies the tail-area based FNDR that is the proportion of non-null p-values among p-values larger than x. On the p-value scale it is defined as

$$\text{Fndr}(x) = \Pr(\text{"alternative"}|X \geq x) = (1 - \eta_0)\frac{1 - F_A(x)}{1 - F(x)}.$$

Fndr and Fdr play a similar role as sensitivity and specificity in classical testing (Genovese and Wassermann, 2002).

Local FDR (fdr) is a density-based quantity defined as the probability of the null under the observed data,

$$\text{fdr}(x) = \Pr(\text{"null"}|X = x) = \frac{\eta_0 f_0(x)}{f(x)} = \frac{\eta_0}{f(x)} \tag{3}$$

with $f(x) = \eta_0 f_0(x) + (1 - \eta_0)f_A(x)$. As with Fdr, to ensure that the local FDR is increasing with x the density $f_A(x)$ is assumed to be monotonically decreasing. The local FNDR is the probability of the alternative under the observed data, and is thus is given by

$$\text{fndr}(x) = 1 - \text{fdr}(x).$$

There is also a direct relationship between fdr and Fdr. As can be seen from its definition Fdr is a conditional average of fdr. Hence, for monotonic $f_A(x)$ we find $\text{Fdr}(x) \leq \text{fdr}(x)$ for all x (Efron, 2008).

Like the HC objective function, fdr and Fdr are scalars and thus transform under a change of coordinates from x to y as $\text{fdr}(y) = \text{fdr}(x(y))$ and $\text{Fdr}(y) = \text{Fdr}(x(y))$.

3.2 Signal identification with FDR and FNDR

A standard approach to obtain a decision threshold with FDR is to refer to the rule of Benjamini and Hochberg (1995) with a cutoff such as $\widehat{\text{Fdr}}(p_{(i)}) \leq 0.05$. Alternatively, a threshold may be found by controlling local FDR, for instance by requiring $\widehat{\text{fdr}}(p_{(i)}) \leq 0.2$ (e.g. Efron, 2008). This ensures that the identified features are mostly from the alternative with only little contamination by unwanted null features. Conversely, if the interest is to identify true null features then similar thresholds may be imposed on FNDR rather than FDR (Ahdesmäki and Strimmer, 2010).

This illustrated for local FDR and local FNDR in Fig. **1b** where the signal space is divided by the decision thresholds x^{fdr} and x^{fndr} into three distinct zones corresponding

to areas where one is very sure about membership to the null (local FNDR < 0.2 or local FDR > 0.8) or to the alternative (local FDR < 0.2) and one additional intermediate region.

From a classification perspective there exists another threshold — the class boundary x^{CB} — that provides a natural separation between null and non-null components. At x^{CB} the probabilities of membership to the alternative and to the null are equal to $1/2$. Hence, in terms of local FDR we have

$$\text{fdr}(x^{\text{CB}}) = \text{fndr}(x^{\text{CB}}) = \frac{1}{2}.$$

As can be seen in Fig. **1b** by construction x^{CB} is located inbetween x^{fndr} and x^{fdr}. From the definition $\text{fdr}(x^{\text{CB}}) = 1/2$ and Eq. **3** we obtain the condition

$$\eta_0 f_0(x^{\text{CB}}) = (1 - \eta_0) f_A(x^{\text{CB}}) \tag{4}$$

for the CB threshold.

4 Comparison of CB and HC decision thresholds

It is now instructive to study the mutual connections among the various decision thresholds, in particular among x^{HC}, x^{KS}, and x^{CB}.

4.1 Kolmogorov-Smirnov (KS) decision threshold

The location x^{KS} where the Kolmogorov-Smirnov objective function $|F_A(x) - F_0(x)|$ is maximized is given by

$$f_0(x^{\text{KS}}) = f_A(x^{\text{KS}}). \tag{5}$$

Thus, the KS decision threshold coincides with the class boundary x^{CB} if $\eta_0 = 1/2$. Thus, the KS threshold implicitly assumes that null and non-null components have the same prior probability.

4.2 HC decision threshold

Using Eq. **2** we may determine the population decision threshold that one tries to estimate by maximizing the empirical HC objective $\widehat{HC}(x)$. This leads to the general condition

$$\begin{aligned} f_0 \left\{ 2F(1-F) + (F_A - F_0)(1 - 2F)\, \eta_0 \right\} = \\ f_A \left\{ 2F(1-F) - (F_A - F_0)(1 - 2F)\, (1 - \eta_0) \right\} \end{aligned} \tag{6}$$

that must be satisfied by the HC decision threshold x^{HC} (note that in Eq. **6** the arguments to F, F_0 and F_A have been left out for the sake of clarity).

There are two cases when the HC threshold condition simplifies substantially. First, if the null and alternative components are well separated: then $F_A(x^{\text{HC}}) = 1$ and $F_0(x^{\text{HC}}) = 0$ and consequently $F(x^{\text{HC}}) = 1 - \eta_0$ so that Eq. **6** reduces to

$$\eta_0 f_0(x^{\text{HC}}) = (1 - \eta_0) f_A(x^{\text{HC}}).$$

Thus, for well-separated null and alternative the HC threshold is identical to the CB threshold.

Second, if null and alternative components are very close: then $F_A(x^{\text{HC}}) \approx F_0(x^{\text{HC}})$ and Eq. **6** becomes

$$f_0(x^{\text{HC}}) = f_A(x^{\text{HC}}),$$

i.e. the HC threshold becomes identical to the KS threshold.

Hence, the HC threshold may be viewed as a compromise between the CB threshold and the KS threshold. This is directly observed in the study of the "rare-weak" model (cf. Tab. **2a**).

5 Rare weak model

The use of "Higher Criticism" is particularly advocated in settings where the signal is sparse and weak. This situation is described by the so-called "rare weak" (RW) model that has been used to study the performance of HC. In the following we introduce the RW model and compare corresponding decision thresholds.

5.1 Setup of RW model

The RW model is a sparse normal mean mixture model with

$$Z \sim (1 - \epsilon) N(0, 1) + \epsilon N(\tau, 1). \tag{7}$$

Its two parameters $\tau \in [0; \infty]$ and $\epsilon \in [0; 1]$ describe intensity and sparsity of the signal. If ϵ is small then the non-null features are rare, and likewise if τ is small then the effect size is weak (hence the name of the model). From this mixture we observe z-scores $z_1, \ldots, z_d$, which provide the data from which decision thresholds are inferred.

Despite its simplicity, this model is sufficiently rich to study the behavior of signal detection and signal identification methods (Ingster, 1999; Donoho and Jin, 2004, 2008, 2009; Xie et al., 2011; Ji and Jin, 2012). A generalized RW model with an additional variance parameter in the alternative is discussed in Cai et al. (2011).

A typical scenario where the RW model naturally arises is in classification. For example consider a two class setting with means $\mu_1 = \mu$ and $\mu_2 = -\mu$ where $\mu = (\ldots, \mu_0, \ldots, 0, \ldots)^T$ is a d-dimensional vector containing either 0 or μ_0 as components and with ϵ describing the proportion of non-zero entries. Further assume an identity covariance I_d and equal number of observations $n_1 = n_2 = n/2$ from the two classes. Then the corresponding z-score vector $(1/n_1 + 1/n_2)^{-1/2}(\hat{\mu}_1 - \hat{\mu}_2)$ used for variable selection (e.g., Zuber and Strimmer, 2009) simplifies to $z = \sqrt{n}\hat{\mu}$. The d components of z

9

follow the RW model of Eq. **7** with $\tau = \sqrt{n}\mu_0$. Note the confounding of n and μ_0, so a small number of observations n and large μ_0 gives rise to the same RW model as large sample size and small μ_0.

Instead of ϵ and τ it is sometimes convenient to use the alternative parameterization

$$\beta_\epsilon = -\log(\epsilon)/\log(d)$$

and

$$r_\tau = \left(\frac{\tau^2}{2}\right)/\log(d)$$

with corresponding backtransformations $\epsilon_\beta = d^{-\beta}$ and $\tau_r = \sqrt{2r\log(d)}$.

The motivation to use β instead of ϵ to measure sparsity is that for d observations the smallest possible fraction of the alternative is $1/d$. The change of variables maps $\epsilon \in [\frac{1}{d}; 1]$ to $\beta \in [0; 1]$. A sparse setting in the RW model is characterized by $\beta \in [\frac{1}{2}, 1]$ or equivalently $\epsilon < d^{-1/2}$. Similarly, the alternative intensity parameter is a map of $\tau \in [0; \sqrt{2\log(d)}]$ to $r \in [0; 1]$. As for d observed z-scores their maximum is bounded in expectation by $\sqrt{2\log(d)}$, a RW model with $r > 1$ contains comparatively well-separated null and alternative components whereas in a model with $r < 1$ the signal is weak.

5.2 Decision boundaries for the RW model

The RW model is simple enough to allow analytical calculations of some decision boundaries.

Using the null and alternative densities $f_0(z) = \frac{1}{\sqrt{2\pi}}e^{-z^2/2}$ and $f_A(z) = \frac{1}{\sqrt{2\pi}}e^{-(z-\tau)^2/2}$ and distribution functions $F_0(z) = \Phi(z)$ and $F_A(z) = \Phi(z-\tau)$ the KS decision threshold (Eq. **5**) for the RW model is

$$z^{\text{KS}} = \frac{\tau}{2}.$$

Similarly, the classification class boundary (Eq. **4**) simplifies for the RW model to

$$z^{\text{CB}} = \frac{\tau}{2} + \frac{1}{\tau}\log\left(\frac{1-\epsilon}{\epsilon}\right).$$

For $\epsilon = 1/2$ the CB threshold reduces to the KS threshold and for $\epsilon \leq 1/2$ we have $z^{\text{CB}} \geq z^{\text{KS}}$. For fixed ϵ and the effect size τ large enough the second term above also vanishes and hence also leads to the KS threshold. As the proportion of non-null features becomes smaller ($\epsilon \to 0$) the decision threshold moves to infinity ($z^{\text{CB}} \to \infty$). Thus, if $\epsilon = 0$ no feature will be classified as non-null.

For the HC decision threshold unfortunately no analytic expression for z^{HC} is available. From the general considerations above (cf. Section 4.2) we know that for larger τ the HC threshold approximates the CB threshold, and that both reduce to the KS threshold for $\epsilon = 1/2$. Furthermore, Donoho and Jin (2009, Appendix Eq. 1.1) show that for the RW model $\text{fdr}(z^{\text{HC}}) \geq 1/2$. This together with the monotonicity of the local FDR in the RW model implies that

$$z^{\text{HC}} \leq z^{\text{CB}}.$$

Thus, in general using the HC decision threshold causes the inclusion of more features than using the CB threshold.

Of particular interest is the behavior of the HC threshold for small values of ϵ. Specifically, if $\epsilon = 0$ and τ is finite then the HC threshold is also finite. For example, $\epsilon = 0$ and $\tau = 2$ leads to $z^{HC} \approx 3.35$, which is distinctly different from the natural class boundary $z^{CB} \to \infty$. Thus, by construction the HC criterion (and also the KS threshold) encourages false positives in signal identification.

A comparison of the KS, HC, and CB thresholds for some settings of ϵ and τ is given in Tab. **2a**. As expected, with increasing τ the HC and CB thresholds become very similar and for $\epsilon = 1/2$ both HC and the CB threshold reduce to the KS threshold. Thus, the pattern confirms the general relationships of these decision thresholds discussed above.

In addition, in the RW model there exist a further close link between the HC and CB thresholds. This results from the special structure of the parameter space of the RW model discussed next.

5.3   Phase space of the RW model

Within the RW model the behavior of signal detection and identification procedures have been studied extensively. This has lead to the remarkable insight that there exist several fundamental boundaries in its phase space that give rise to four distinct regions, as illustrated in Fig. **2a**.

Ingster (1999) discovered the *detection boundary*

$$r_{\text{detect}}(\beta) = \begin{cases} \beta - \frac{1}{2} & \beta \in [\frac{1}{2}; \frac{3}{4}] \\ (1 - \sqrt{1 - \beta})^2 & \beta \in [\frac{3}{4}; 1]. \end{cases}$$

Below this boundary lies the "undetectable" region in which even signal detection is impossible, i.e. no method is able to decide whether $\epsilon \neq 0$. Conversely, above the detection boundary it is possible to consistently estimate ϵ (Cai et al., 2007).

Donoho and Jin (2004) report the *identification boundary*

$$r_{\text{ident}}(\beta) = \beta.$$

It is only above this boundary in the "estimable" and "recoverable" regions that signal identification by thresholding is actually possible. In terms of original parameters this corresponds to the conditions $\tau \geq \sqrt{-2 \log(\epsilon)}$ or $\epsilon \geq \exp(-\tau^2/2)$. Directly below this boundary lies the "detectable" region where detection of a signal is possible but not identification. This shows that signal identification is more difficult than signal detection.

Finally, Xie et al. (2011) and Ji and Jin (2012) demonstrated the existence of the *recovery boundary*

$$r_{\text{recov}}(\beta) = (1 + \sqrt{1 - \beta})^2$$

above which in the "recoverable" region almost all signal can be completely identified.

Table 2: a) Comparison of the KS, HC and CB decision thresholds in the RW model, and b) Analysis of four cancer gene expression data sets with shrinkage discriminant analysis.

a) Comparison of Thresholds

Setting	z^{KS}	z^{HC}	z^{CB}
$\tau = 2$			
$\epsilon = 0$	1	3.3514	∞
$\epsilon = 0.001$	1	3.0707	4.4534
$\epsilon = 0.01$	1	2.5203	3.2976
$\epsilon = 0.1$	1	1.7574	2.0986
$\epsilon = 0.5^*$	1	1.0000	1
$\tau = 4$			
$\epsilon = 0$	2	3.3514	∞
$\epsilon = 0.001^*$	2	3.6377	3.7267
$\epsilon = 0.01^*$	2	3.0965	3.1488
$\epsilon = 0.1^*$	2	2.5268	2.5493
$\epsilon = 0.5^*$	2	2.0000	2
$\tau = 6$			
$\epsilon = 0$	3	8.1607	∞
$\epsilon = 0.001^*$	3	4.1454	4.1511
$\epsilon = 0.01^*$	3	3.7631	3.7659
$\epsilon = 0.1^*$	3	3.3652	3.3662
$\epsilon = 0.5^*$	3	3.0000	3

* Signal identification is possible as $\epsilon \geq \exp(-\tau^2/2)$, see Section 5.3.

b) Cancer Gene Expression Data

Data / Method	Prediction Error		Selected Variables
Prostate ($d = 6033, n = 102, K = 2$)			
CB	0.0637	(0.0053)	115
HC	0.0497	(0.0045)	116
FNDR	0.0550	(0.0048)	131
Lymphoma ($d = 4026, n = 62, K = 3$)			
CB	0.0211	(0.0042)	178
HC	0.0000	(0.0000)	345
FNDR	0.0036	(0.0018)	392
SRBCT ($d = 2308, n = 63, K = 4$)			
CB	0.0000	(0.0000)	88
HC	0.0007	(0.0007)	174
FNDR	0.0000	(0.0000)	89
Brain ($d = 5597, n = 42, K = 5$)			
CB	0.1633	(0.0120)	78
HC	0.1417	(0.0108)	131
FNDR	0.1525	(0.0120)	102

K: number of classes in the response variable.

a) Phase Space of Rare–Weak Model

b) Ratio of HC and CB Thresholds at the Signal Identification Boundary and Above
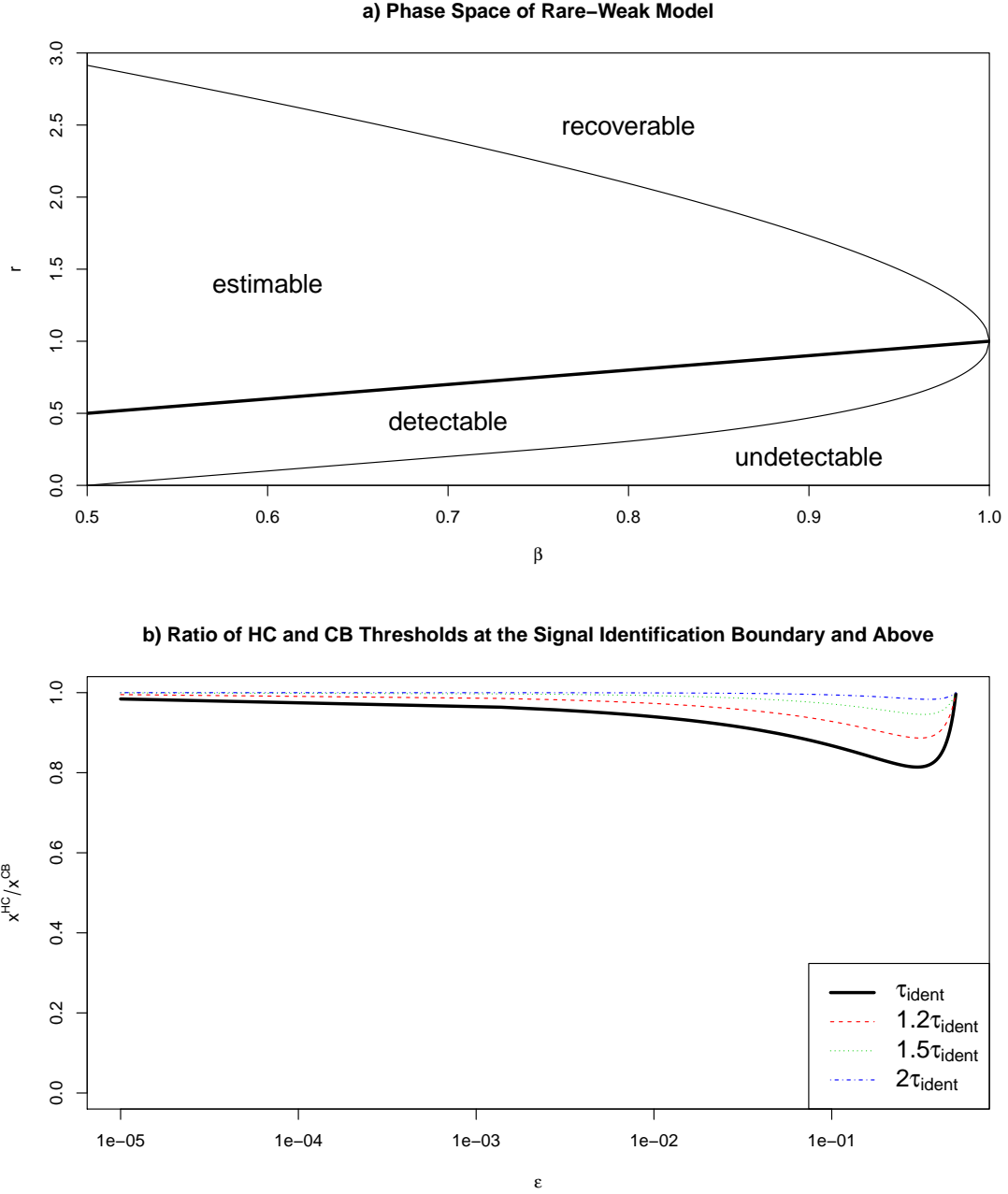
Figure 2: a) Phase space of the RW model following Xie et al. (2011) and Ji and Jin (2012). The bold line shows the signal identification boundary $r_{\mathrm{ident}}(\beta) = \beta$ above which signal identification is possible. For details on the four regions see the description in Section 5.3. b) Ratio of x^{HC} and x^{CB} thresholds at the signal identification boundary (solid line) and above (dotted lines). Note that $\tau_{\mathrm{ident}}(\epsilon) = \sqrt{-2\log(\epsilon)}$.

5.4 HC threshold as approximation of the natural class boundary

When comparing the KS, HC and CB decision thresholds in Tab. **2a** a striking phenomenon can be observed: whenever signal identification is possible, i.e. if $\epsilon \geq \exp(-\tau^2/2)$, then then z^{CB} and z^{HC} are very similar.

To investigate this further we computed the ratio of the HC and CB threshold directly at the signal identification boundary, and above (Fig. **2b**). Already at the boundary this ratio is close to 1, especially for small values of ϵ. Moving further into the "estimable" and "recoverable" regions the differences between the two thresholds become negligible.

Hence, in the RW model in the area where signal identification is possible z^{HC} and z^{CB} are in the worst case very similar and mostly indistinguishable for practical purposes.

6 Data examples

To further study the relationship among the HC, CB, and FNDR decision thresholds we analyzed both simulated as well as experimental data.

6.1 Synthetic data

We simulated data from the RW model at the signal identification boundary and above, as follows:

1. We sampled $d = 10,000$ z-scores from the mixture model Eq. **7** with $\epsilon = 0.01$ and $\tau \in \{3, 4, 5, 6\}$. For $\tau = 3$ this is is a sparse and weak scenario located directly at the signal identification boundary ($\epsilon \approx \exp(-\tau^2/2)$).

2. From the test statistics $z_1, \ldots, z_d$ we computed p-values according to $p_i = 1 - F_0(z_i)$.

3. Subsequently, the empirical HC threshold was obtained by maximization of Eq. **1**.

4. In addition, local FDR was estimated using the `fdrtool` algorithm (Strimmer, 2008a,b) and correspondingly the CB (local FDR = 0.5) and FNDR (local FDR = 0.8) decision thresholds were identified.

5. For each of the three investigated thresholds (HC, CB, FNDR) the number of false positives (FP), false negatives (FN), true positives (TP) and true negatives (TN) were determined.

6. The simulations were repeated $B = 1000$ times to estimate mean errors and their standard deviations.

The results are visualized in Fig. **3**. As expected, the HC and CB thresholds yield similar results with growing τ. However, if the signal is weak (small τ) signal identification with HC leads to many more more false positives, and in addition the variability
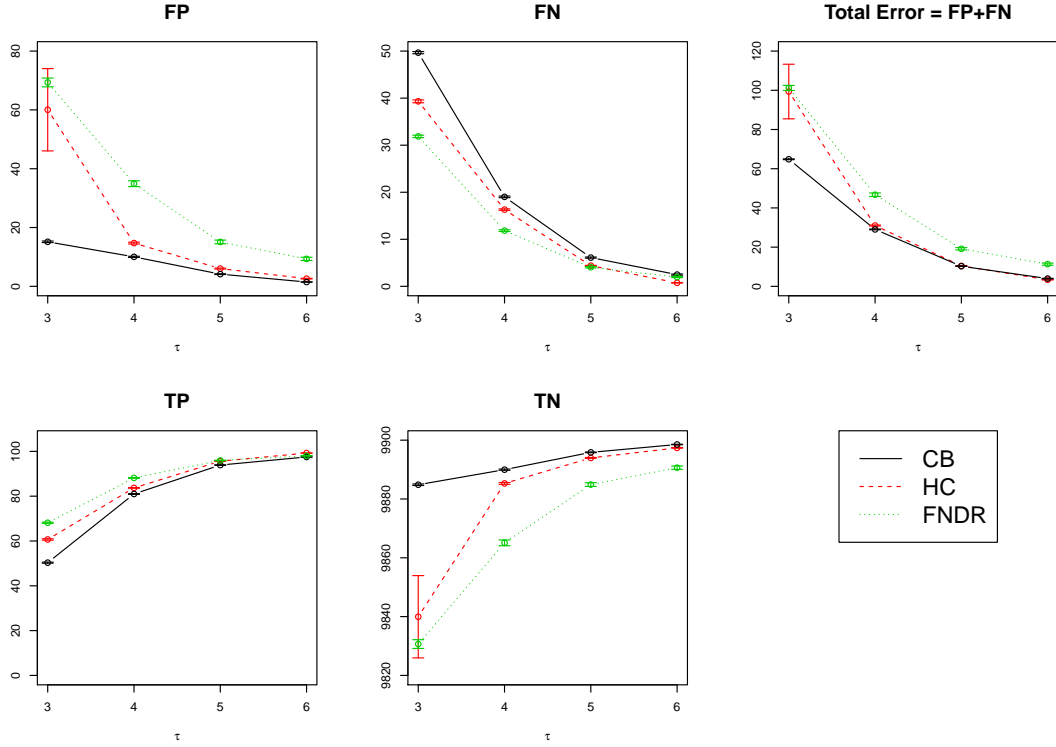
Figure 3: Comparison of errors when using the HC, CB, and FNDR decision thresholds on data simulated from the RW model located directly at the detection boundary ($\epsilon = 0.01$ and $\tau = 3$) and above ($\tau > 3$).

of the error rates for HC is very large. Conversely, in this situation the CB threshold is more cautious and thus results in more false negatives. For all settings the error rates of HC are found in between those of CB and FNDR. Interestingly, the total error (FP+FN) is smallest when using the CB threshold.

We also repeated this study with other sparsity settings $\epsilon > 0.01$. The resulting error plots all show exactly the same pattern of convergence of the CB and and HC methods as Fig. **3**.

6.2   Gene expression data

Next, we also analyzed four clinical gene expression data sets related to prostate cancer (Singh et al., 2002) lymphoma (Alizadeh et al., 2000), small round blue cell tumors (SRBCT) (Khan et al., 2001), and brain cancer (Pomeroy et al., 2002). Previously, in Ahdesmäki and Strimmer (2010) we have compared the relative effectiveness of the FNDR and HC thresholds to select relevant genes in shrinkage discriminant analysis

using CAT scores (Zuber and Strimmer, 2009).

In Tab. **2b** we show in addition the estimated prediction error and the number of selected variables for the CB threshold. Generally, using the CB decision threshold leads to the smallest predictor sets. Except for the prostate data the number of selected genes is roughly half compared to using the HC threshold as criterion. As the predictor error is only slightly increased we conclude that most of the additionally included predictors by HC are false positives.

For practical analysis of gene expression data this implies that using x^{CB} yields — in comparison with x^{HC} — smaller and hence more interpretable predictor gene sets without compromising prediction error.

7 Discussion

Our investigation of the relationship of the HC and FDR methods started with the aim to better understand HC as a method for signal identification. In the context of variable selection for classification Donoho and Jin (2008) demonstrated empirically that using x^{HC} as a decision threshold outperforms competing procedures, in particular those using a threshold based on FDR. Donoho and Jin (2009) further justified HC as a signal identification procedure by showing that x^{HC} minimizes an approximation to the missclassification error.

Here, we argue that the HC decision threshold may also be viewed as an approximation of the natural class boundary between the null and alternative groups in the RW mixture model. This CB threshold can be directly expressed in terms of local FDR and local FNDR. Importantly, in the RW model in the region of the phase space where signal identification is possible both thresholds are either very similar or practically indistinguishable. Interestingly, computing this threshold via HC uses only distribution functions (F, F_0, and F_A, cf. Eq. **2**) but in addition requires optimization, whereas computation via local FDR is direct but employs densities (f, f_0, and f_A, cf. Eq. **3**) which are more difficult to obtain.

If the two thresholds are notably different then using the HC threshold leads to the inclusion of more false positives, and conversely the CB threshold yields a more compact feature set but with slightly increased prediction error. In short, the CB threshold is more cautious than the HC threshold (and the FNDR threshold).

Hence, our study provides further support to the excellent performance of HC for signal identification. However, our conclusions are different from that of Donoho and Jin (2008, 2009). First, we show that false discovery rates, properly applied, are indeed perfectly useful for signal identification, which has been disputed earlier. Second, the convergence of the CB and HC thresholds in the "estimable" and "recoverable" regions indicates that this is what HC is actually approximating.

In general, estimation of the CB threshold is a challenging problem as this requires the fit of a mixture model and estimation of the mixing density. In contrast, the empirical HC threshold can readily be determined using p-values computed from F_0 alone. Thus,

for signal identification the HC approach provides a simple yet effective means to approximate the CB threshold.

Acknowledgements

References

Ahdesmäki, M. and Strimmer, K. (2010). Feature selection in omics prediction problems using cat scores and false non-discovery rate control. *Ann. Appl. Statist.*, 4:503–519.

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O., and Staudt, L. M. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403:503–511.

Anderson, T. W. and Darling, D. A. (1954). A test of goodness of fit. *J. Amer. Statist. Assoc.*, 49:765–769.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*, 57:289–300.

Cai, T. T., Jeng, X. J., and Jin, J. (2011). Optimal detection of heterogeneous and heteroscedastic mixtures. *J. R. Statist. Soc. B*, 73:629–662.

Cai, T. T., Jin, J., and Low, M. G. (2007). Estimation and confidence sets for spare normal mixtures. *Ann. Statist.*, 35:2421–2449.

Darling, D. A. (1957). The Kolmogorov-Smirnov, Cramér-von Mises tests. *Ann. Math. Stat.*, 28:823–838.

Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Annals of Statistics*, 32:962–994.

Donoho, D. and Jin, J. (2008). Higher criticism thresholding: optimal feature selection when useful features are rare and weak. *Proc. Natl. Acad. Sci. USA*, 105:14790–15795.

Donoho, D. and Jin, J. (2009). Feature selection by higher criticism thresholding achieves the optimal phase diagram. *Phil. Trans. R. Soc. A*, 367:4449–4470.

Efron, B. (2008). Microarrays, empirical Bayes, and the two-groups model. *Statist. Sci.*, 23:1–22.

Genovese, C. and Wassermann, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *J. R. Statist. Soc. B*, 64:499–517.

Ingster, Y. I. (1999). Minimax detection of a signal for l_n^p balls. *Math. Methods. Statist.*, 7:401–428.

Jager, L. and Wellner, J. A. (2007). Goodness-of-fit tests via phi-divergences. *Ann. Statist.*, 35:2018–2053.

Ji, P. and Jin, J. (2012). UPS delivers optimal phase diagram in high-dimensional variable selection. *Ann. Statist.*, 40:73–103.

Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C., and Meltzer, P. S. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Med.*, 7:673–679.

Langaas, M., Lindqvist, B. H., and Ferkingstad, E. (2005). Estimating the proportion of true null hypotheses, with application to DNA microarray data. *J. R. Statist. Soc. B*, 67:565–572.

Pomeroy, S. L., Tamayo, P., Gaasenbeek, M., Sturla, L. M., Angelo, M., McLaughlin, M. E., Kim, J. Y. H., Goumnerova, L. C., Black, P. M., Lau, C., Allen, J. C., Zagzag, D., Olson, J. M., Curran, T., Wetmore, C., Biegel, J. A., Poggio, T., Mukherjee, S., Rifkin, R., Califano, A., Stolovitzky, G., Louis, D. N., Mesirov, J. P., Lander, E. S., and Golub, T. R. (2002). Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 415:436–442.

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A. V., Richie, J. P., Lander, E. S., Loda, M., Kantoff, P. W., Golub, T. R., and Sellers, W. R. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1:203–209.

Strimmer, K. (2008a). fdrtool: a versatile R package for estimating local and tail area-based false discovery rates. *Bioinformatics*, 24:1461–1462.

Strimmer, K. (2008b). A unified approach to false discovery rate estimation. *BMC Bioinformatics*, 9:303.

Tukey, J. W. (1976). T13 N: the higher criticism. Course Notes, Statistics 411, Princeton Univ.

Xie, J., Cai, T. T., and Li, H. (2011). Sample size and power analysis for sparse signal recovery in genome-wide association studies. *Biometrika*, 98:273–290.

Zuber, V. and Strimmer, K. (2009). Gene ranking and biomarker discovery under correlation. *Bioinformatics*, 25:2700–2707.